Bias-Variance Tradeoff

Optimal Prediction

Goal: Predict $Y \in \mathbb{R}^d$ **given** $X \in \mathbb{R}^d$ **if** $(X, Y) \sim P_{XY}$

Find function η that minimizes

$$\mathbb{E}_{XY}[(Y - \eta(X))^2] = \mathbb{E}_X\left[\underline{\mathbb{E}_{Y|X}[(Y - \eta(x))^2 | X = x]}\right]$$

(Hint: for any x, $\eta(x) = c_x$ where c_x minimizes $\mathbb{E}_{Y|X}[(Y - c_x)^2 | X = x]$)

Optimal Prediction

Goal: Predict $Y \in \mathbb{R}^d$ **given** $X \in \mathbb{R}^d$ **if** $(X, Y) \sim P_{XY}$

Find function η that minimizes

$$\mathbb{E}_{XY}[(Y - \eta(X))^2] = \mathbb{E}_X\left[\mathbb{E}_{Y|X}[(Y - \eta(x))^2 | X = x]\right]$$

(Hint: for any x, $\eta(x) = c_x$ where c_x minimizes $\mathbb{E}_{Y|X}[(Y - c_x)^2 | X = x]$)

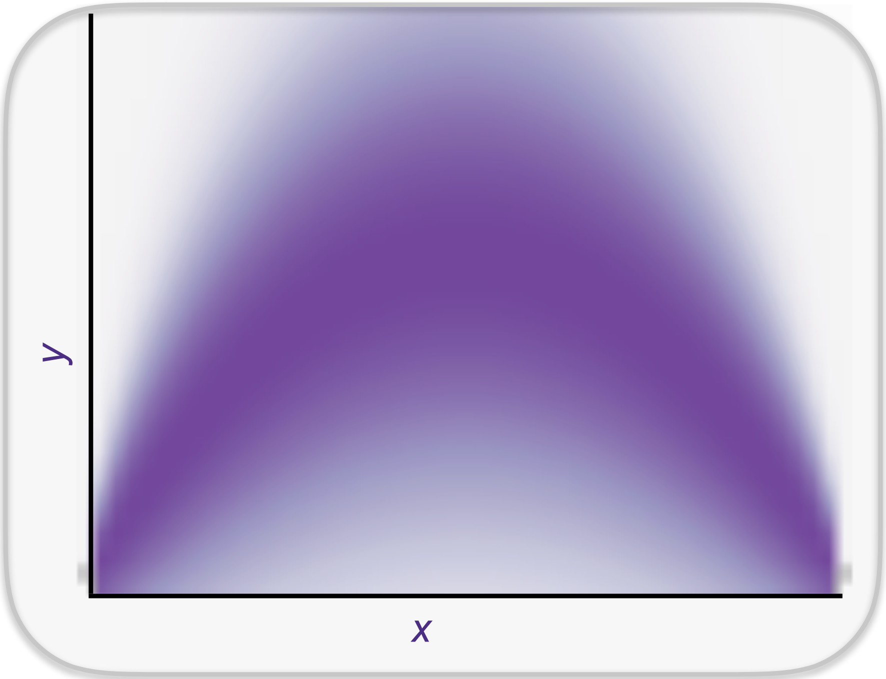
$$\begin{aligned}
0 &= \frac{d}{dc_x}\mathbb{E}_{Y|X}[(Y - c_x)^2 | X = x] \\
&= \mathbb{E}_{Y|X}[\frac{d}{dc_x}(Y - c_x)^2 | X = x] \\
&= \mathbb{E}_{Y|X}[-2(Y - c_x) | X = x] = -2\mathbb{E}_{Y|X}[Y | X = x] + 2c_x
\end{aligned}$$

Squared Error Optimal Predictor: $\eta(x) = \mathbb{E}_{Y|X}[Y | X = x]$

Statistical Learning

$$\mathbb{E}_{XY}[(Y - \eta(X))^2]$$

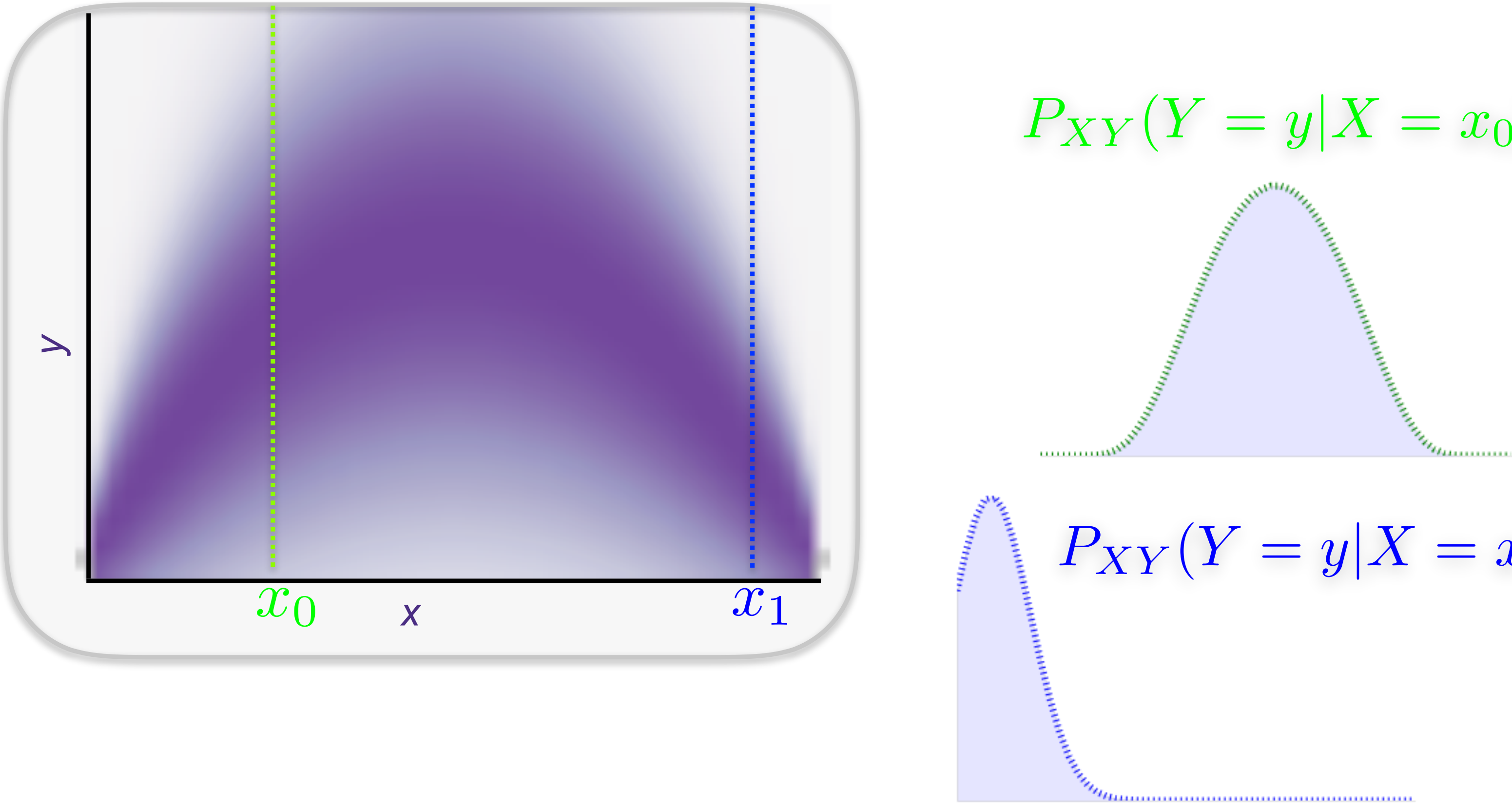
$$P_{XY}(X = x, Y = y)$$



Statistical Learning

$$\mathbb{E}_{XY}[(Y - \eta(X))^2]$$

$P_{XY}(X = x, Y = y)$

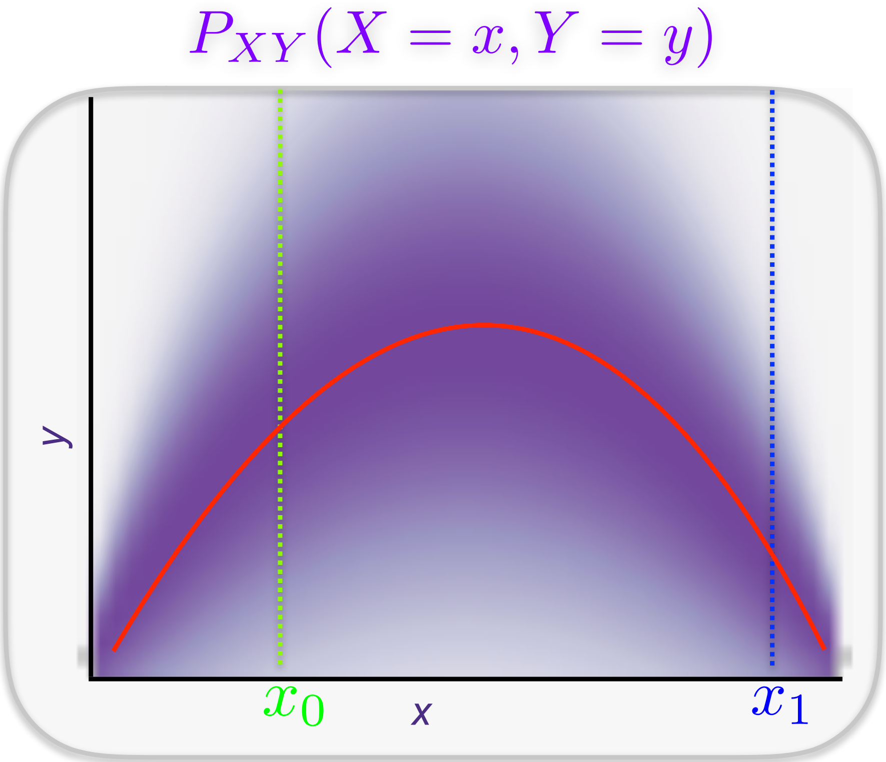


$P_{XY}(Y = y | X = x_0)$

$P_{XY}(Y = y | X = x_1)$

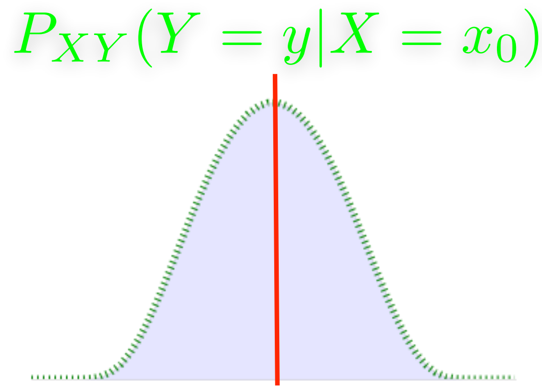
Statistical Learning

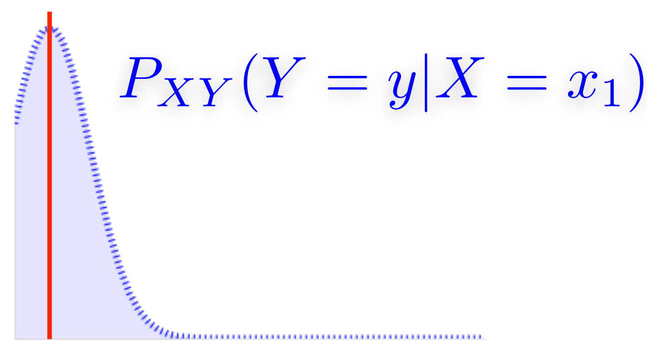
$$\mathbb{E}_{XY}[(Y-\eta(X))^2]$$



Ideally, we want to find:

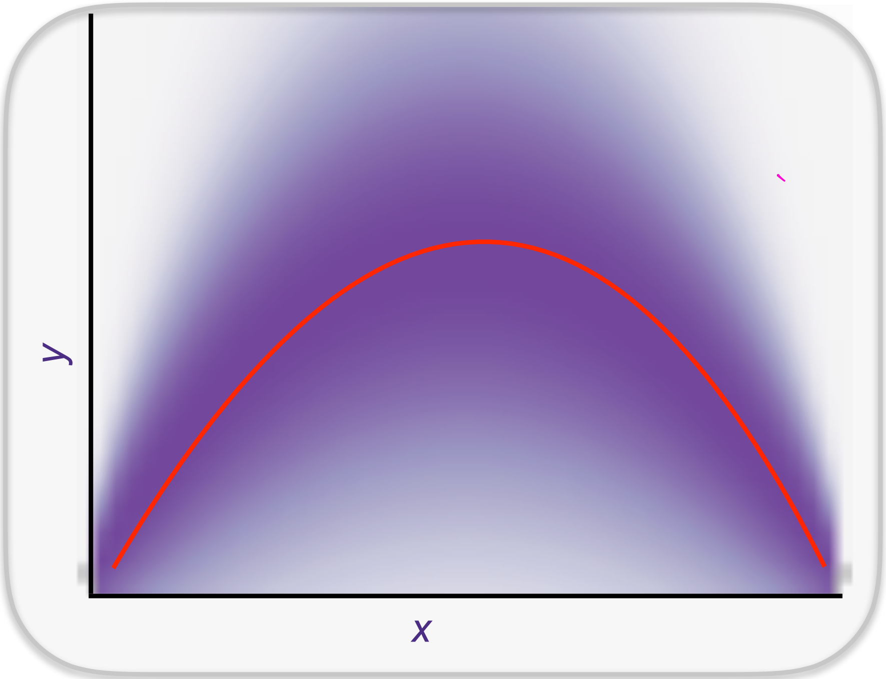
$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$





Statistical Learning

$$P_{XY}(X = x, Y = y)$$

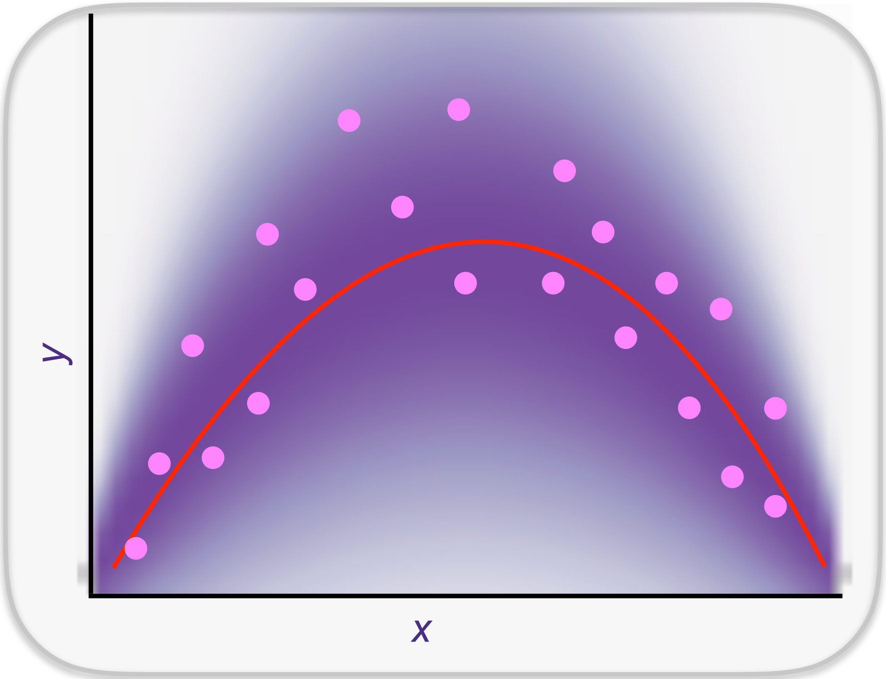


Ideally, we want to find:

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

Statistical Learning

$P_{XY}(X = x, Y = y)$



Ideally, we want to find:

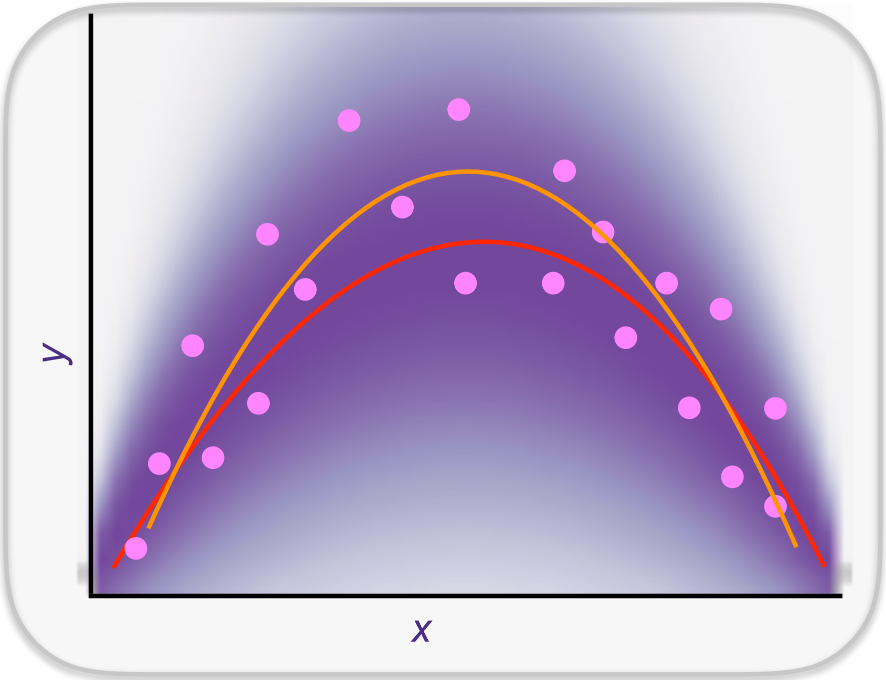
$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

But we only have samples:

$(x_i, y_i) \overset{i.i.d.}{\sim} P_{XY}$ for $i = 1, \ldots, n$

Statistical Learning

$P_{XY}(X = x, Y = y)$



Ideally, we want to find:

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

But we only have samples:

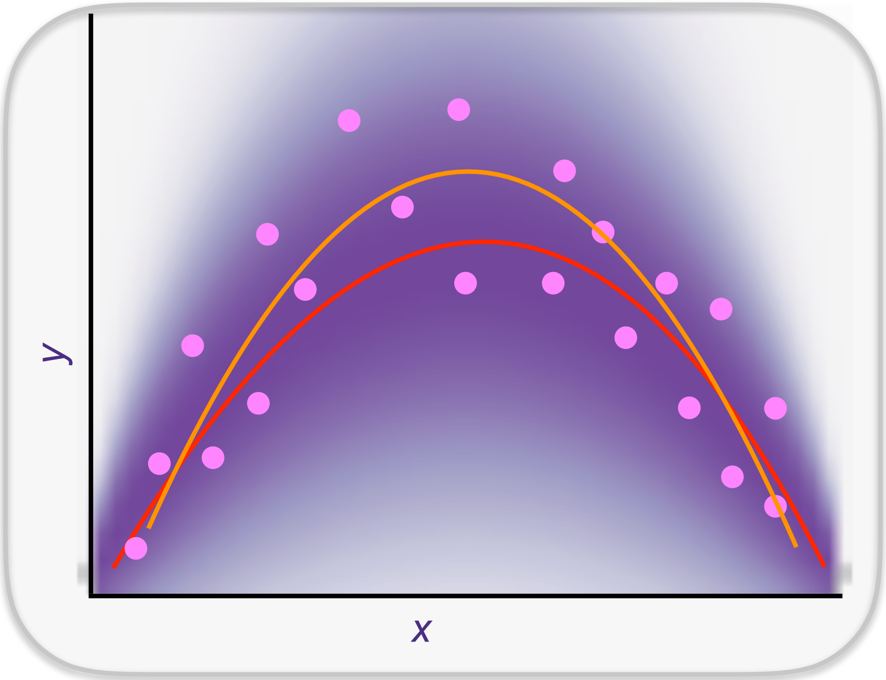
$(x_i, y_i) \overset{i.i.d.}{\sim} P_{XY}$ for $i = 1, \ldots, n$

and are restricted to a function class (e.g., linear) so we compute:

$$\widehat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$$

Statistical Learning

$P_{XY}(X = x, Y = y)$



Ideally, we want to find:

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

But we only have samples:

$(x_i, y_i) \overset{i.i.d.}{\sim} P_{XY}$ for $i = 1, \ldots, n$

and are restricted to a function class (e.g., linear) so we compute:

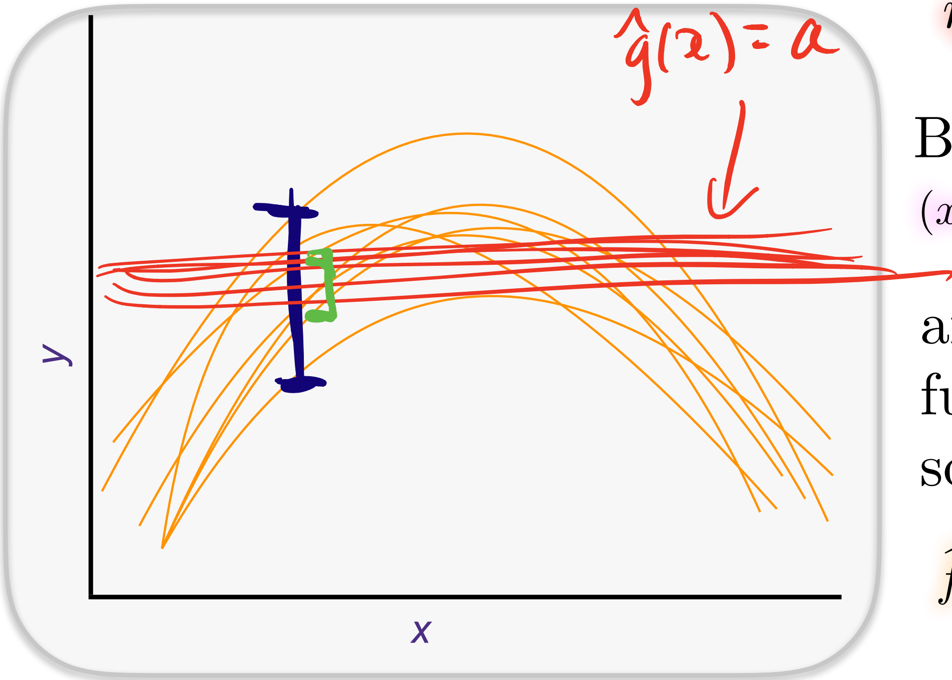
$$\widehat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$$

We care about future predictions: $\mathbb{E}_{XY}[(Y - \widehat{f}(X))^2]$

Statistical Learning

$\hat{f}(x) = a + bx + cx^2$

$P_{XY}(X = x, Y = y)$



Ideally, we want to find:

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

But we only have samples:

$(x_i, y_i) \overset{i.i.d.}{\sim} P_{XY} \quad$ for $i = 1, \ldots, n$

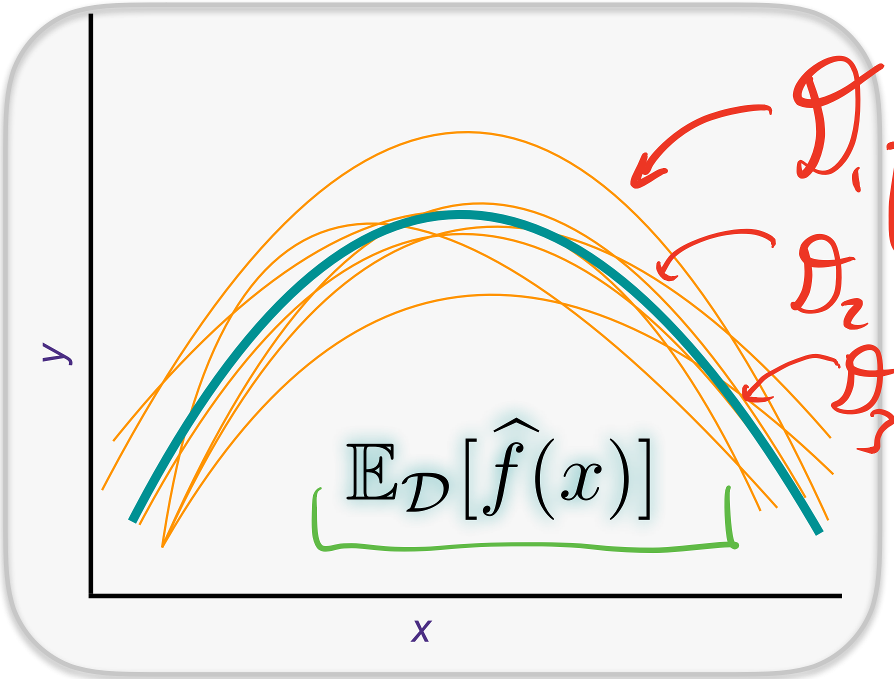
and are restricted to a function class (e.g., linear) so we compute:

$$\widehat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$$

Each draw $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ results in different $\widehat{f}$

Statistical Learning

$P_{XY}(X = x, Y = y)$



Ideally, we want to find:

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

~~But we only have samples:~~

$$\boxed{(x_i, y_i) \overset{i.i.d.}{\sim} P_{XY} \quad \text{for } i = 1, \ldots, n}$$

and are restricted to a function class (e.g., linear) so we compute:

$$\widehat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$$

Each draw $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ results in different $\widehat{f}$

Bias-Variance Tradeoff

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X=x] \qquad \widehat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$$

$$\mathbb{E}_{Y|X}[\mathbb{E}_{\mathcal{D}}[(Y - \widehat{f}_{\mathcal{D}}(x))^2] \big| X = x] = \mathbb{E}_{Y|X}[\mathbb{E}_{\mathcal{D}}[(Y - \eta(x) + \eta(x) - \widehat{f}_{\mathcal{D}}(x))^2] \big| X = x]$$

Bias-Variance Tradeoff

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x] \qquad \widehat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$$

$$\begin{aligned}
\mathbb{E}_{Y|X}[\mathbb{E}_{\mathcal{D}}[(Y - \widehat{f}_{\mathcal{D}}(x))^2] \big| X = x] &= \mathbb{E}_{Y|X}[\mathbb{E}_{\mathcal{D}}[(Y - \eta(x) + \eta(x) - \widehat{f}_{\mathcal{D}}(x))^2] \big| X = x] \\
&= \mathbb{E}_{Y|X}\Big[\mathbb{E}_{\mathcal{D}}[(Y - \eta(x))^2 + 2(Y - \eta(x))(\eta(x) - \widehat{f}_{\mathcal{D}}(x)) \\
&\quad + (\eta(x) - \widehat{f}_{\mathcal{D}}(x))^2] \big| X = x\Big] \\
&= \underbrace{\mathbb{E}_{Y|X}[(Y - \eta(x))^2 \big| X = x]} + \underbrace{\mathbb{E}_{\mathcal{D}}[(\eta(x) - \widehat{f}_{\mathcal{D}}(x))^2]}
\end{aligned}$$

irreducible error
Caused by stochastic label noise

learning error
Caused by either using too "simple" of a model or not enough data to learn the model accurately

Bias-Variance Tradeoff

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

$$\widehat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$$

$$\underline{\mathbb{E}_{\mathcal{D}}[(\eta(x) - \widehat{f}_{\mathcal{D}}(x))^2]} = \mathbb{E}_{\mathcal{D}}[(\eta(x) - \mathbb{E}_{\mathcal{D}}[\widehat{f}_{\mathcal{D}}(x)] + \mathbb{E}_{\mathcal{D}}[\widehat{f}_{\mathcal{D}}(x)] - \widehat{f}_{\mathcal{D}}(x))^2]$$

Bias-Variance Tradeoff

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x] \qquad \widehat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$$

$$\mathbb{E}_{\mathcal{D}}[(\eta(x) - \widehat{f}_{\mathcal{D}}(x))^2] = \mathbb{E}_{\mathcal{D}}[(\eta(x) - \mathbb{E}_{\mathcal{D}}[\widehat{f}_{\mathcal{D}}(x)] + \mathbb{E}_{\mathcal{D}}[\widehat{f}_{\mathcal{D}}(x)] - \widehat{f}_{\mathcal{D}}(x))^2]$$

$$= \mathbb{E}_{\mathcal{D}}[(\eta(x) - \mathbb{E}_{\mathcal{D}}[\widehat{f}_{\mathcal{D}}(x)])^2 + 2(\eta(x) - \mathbb{E}_{\mathcal{D}}[\widehat{f}_{\mathcal{D}}(x)])(\mathbb{E}_{\mathcal{D}}[\widehat{f}_{\mathcal{D}}(x)] - \widehat{f}_{\mathcal{D}}(x))$$

$$+ (\mathbb{E}_{\mathcal{D}}[\widehat{f}_{\mathcal{D}}(x)] - \widehat{f}_{\mathcal{D}}(x))^2]$$

$$= (\eta(x) - \mathbb{E}_{\mathcal{D}}[\widehat{f}_{\mathcal{D}}(x)])^2 + \mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[\widehat{f}_{\mathcal{D}}(x)] - \widehat{f}_{\mathcal{D}}(x))^2]$$

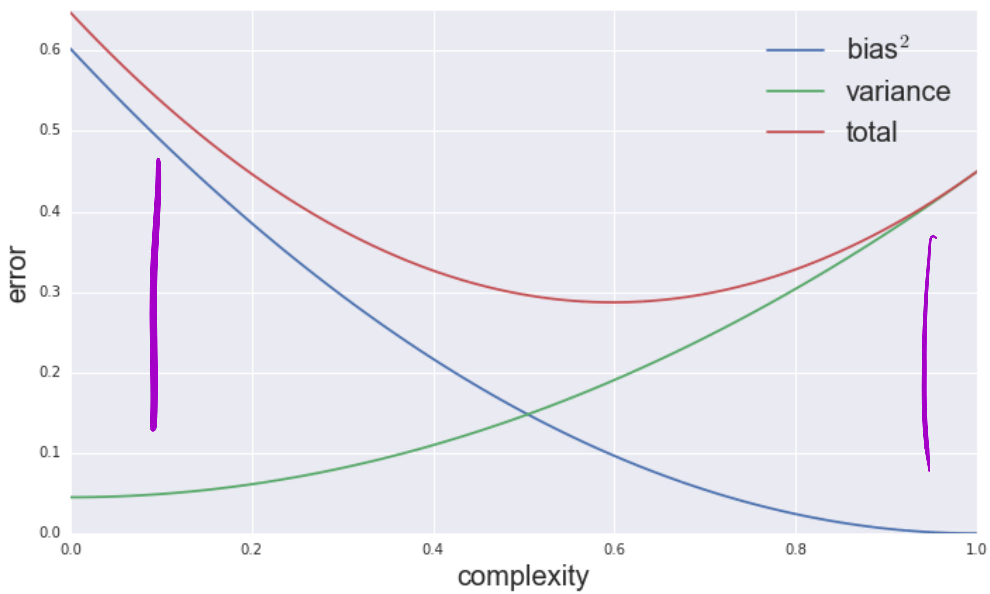
biased squared **variance**

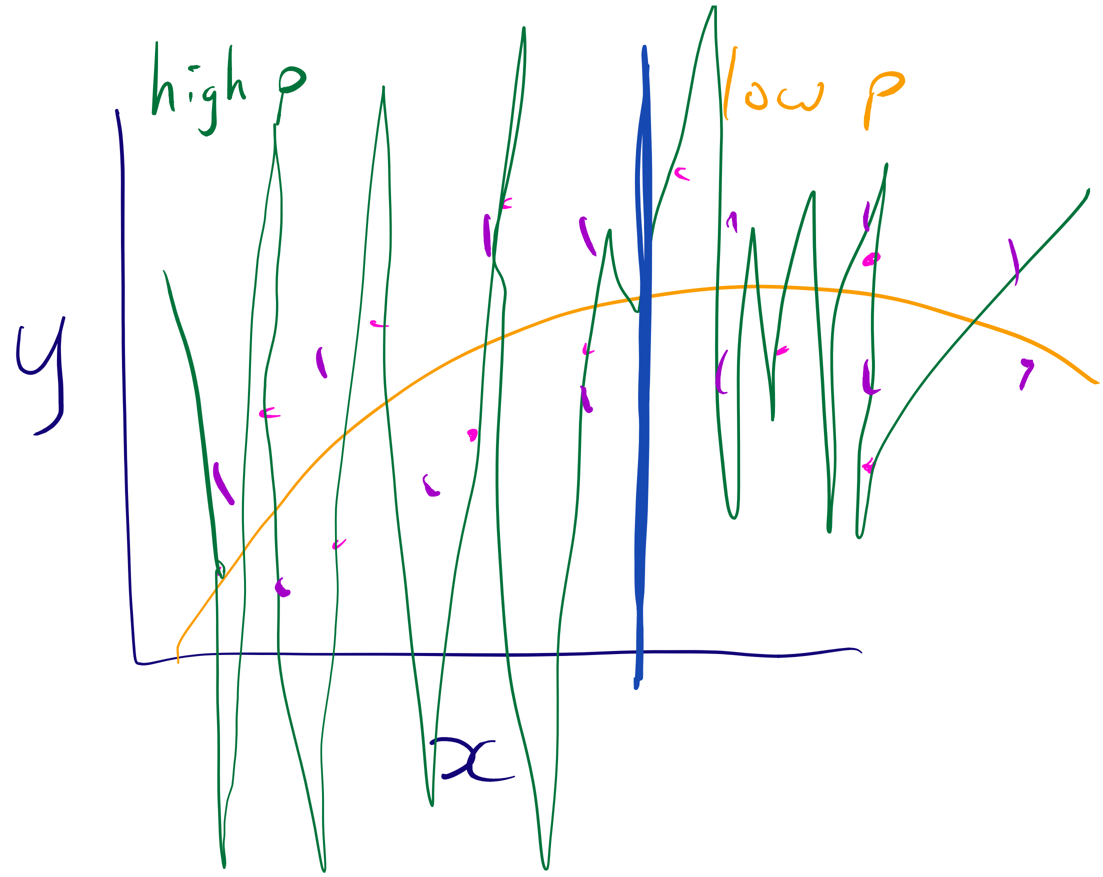
Bias-Variance Tradeoff

$\eta(x)$

$$\mathbb{E}_{Y|X}[\mathbb{E}_{\mathcal{D}}[(Y - \widehat{f}_{\mathcal{D}}(x))^2]|X = x] = \underbrace{\mathbb{E}_{Y|X}[(Y - \eta(x))^2|X = x]}_{\text{irreducible error}}$$

$$+ \underbrace{(\eta(x) - \mathbb{E}_{\mathcal{D}}[\widehat{f}_{\mathcal{D}}(x)])^2}_{\text{biased squared}} + \underbrace{\mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[\widehat{f}_{\mathcal{D}}(x)] - \widehat{f}_{\mathcal{D}}(x))^2]}_{\text{variance}}$$





$$\hat{f}(x) = \sum_{k=0}^{p} \underline{\underline{a_k}}\, x^k$$

$$X = \begin{bmatrix} — x_1^T — \\ — x_2^T — \\ \vdots \\ — x_n^T — \end{bmatrix}$$

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{(column vector in } \mathbb{R}^d\text{)}$$

$$\mu_j = \frac{1}{n}\sum_{i=1}^{n} x_{i,j} \leftarrow \text{jth component of the ith example}$$

"standardize data"

$$\sigma_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{i,j} - \mu_j)^2$$

Overfitting

$$\forall i,j \text{ set } \tilde{x}_{i,j} = (x_{i,j} - \mu_j)/\sigma_j$$

$$\tilde{\mu}_j = \frac{1}{n}\sum_{i=1}^{n}\tilde{x}_{i,j} = 0$$

$$\tilde{\sigma}_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\tilde{x}_{i,j} - \tilde{\mu}_j)^2 = 1$$

Your data $\{(x_i, y_i)\}_{i=1}^{n}$, $\quad x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$

Want to fit $w \in \mathbb{R}^d$, $b \in \mathbb{R}$ s.t. $y_i \approx x_i^T w + b$

$$\hat{w}, \hat{b} = \underset{w,b}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - (w^T x_i + b)\right)^2 = \underset{w,b}{\operatorname{argmin}} \|y - (Xw + b\mathbf{1})\|_2^2$$

$$\mu_j = \frac{1}{n} \sum_{i=1}^{n} x_{i,j}$$

$$\tilde{x}_{i,j} = x_{i,j} - \mu_j$$

$$\bar{y} = \frac{1}{n} \sum_{k=1}^{n} y_k$$

$$\tilde{y}_i = y_i - \bar{y}$$

$\Rightarrow$ $\tilde{x}, \tilde{y}$ are both mean 0.

$$\Rightarrow \left[\begin{array}{l} \tilde{X}^T \tilde{X} \tilde{w} + \hat{b}\, \cancel{\tilde{X}^T \mathbf{1}}^{\,0} = \tilde{X}^T \tilde{y} \\ \cancel{\mathbf{1}^T \tilde{X}}^{\,0}\, \tilde{b} + \tilde{b}\, \underbrace{\mathbf{1}^T \mathbf{1}}_{n} = \mathbf{1}^T \tilde{y} \end{array} \right.$$

$$\tilde{w} = (\tilde{X}^T \tilde{X})^{-1} \tilde{X} \tilde{y}$$

$$\tilde{b} = \frac{1}{n} \mathbf{1}^T \tilde{y} = 0$$

$$\tilde{y}_i \approx \tilde{x}_i^T \hat{w}$$

$$\tilde{y}_i = y_i - \bar{y} \approx (x_i - \mu)^T \hat{w}$$

$$\Rightarrow \quad y_i \approx (x_i - \mu)^T \hat{w} + \bar{y}$$

$$= x_i^T \hat{w} - \mu^T \hat{w} + \bar{y}$$

$$= \underline{x_i^T \tilde{w} + (\bar{y} - \mu^T \tilde{w})}$$

So let $\hat{w} = \tilde{w}$, $\quad \hat{b} = \bar{y} - \mu^T \tilde{w}$

$$y_i \approx x_i^T \hat{w} + \hat{b}$$

$$\hat{w}, \hat{b} = \underset{w,b}{\arg\min} \sum_{i=1}^{n} (y_i - (w^T x_i + b))^2$$

$$= \underset{w,b}{\arg\min} \underbrace{\|y - (Xw + b\mathbb{1})\|_2^2}$$

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} - x_1^T - \\ \vdots \\ - x_n^T - \end{bmatrix}$$

$$\nabla_w(\cdot) = 2 X^T(y - (Xw + b\mathbb{1}))$$

$$= 2\left(X^T y - X^T X w - b X^T \mathbb{1}\right) = 0$$

$$X^T X w + b X^T \mathbb{1} = X^T y$$

$$\nabla_b(\cdot) = \rightarrow$$

$\{(x_i, y_i)\}_{i=1}^{n}$ Fit w, b s.t. $y_i \approx x_i^T w + b$

$$x_i^T w + b = [x_{i,1}, \ldots, x_{i,d}]^T \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix} + 1 \cdot b$$

$$\vec{X} = \begin{bmatrix} X, & \mathbb{1} \end{bmatrix} \qquad = \underbrace{[x_{i,1}, \ldots, x_{i,d}, 1]}_{\vec{x}_i} \underbrace{\begin{bmatrix} w_1 \\ \vdots \\ w_d \\ b \end{bmatrix}}_{\vec{w}}$$

$$\|y - \vec{X}\vec{w}\|_2^2 \equiv \|y - (Xw + b\mathbb{1})\|_2^2 \qquad =: \vec{x}_i^T \vec{w}$$

$x_i \rightarrow \vec{x}_i$ by appending a 1 to the end

$w \rightarrow \vec{w}$ by inflating dim by 1 $\quad \sum_{i=1}^{n} (y_i - \vec{w}^T \vec{x}_i)^2$

Bias-Variance Tradeoff

- Choice of hypothesis class introduces learning bias
  - More complex class → less bias
  - More complex class → more variance
- But in practice??

Bias-Variance Tradeoff

- Choice of hypothesis class introduces learning bias
  - More complex class → less bias
  - More complex class → more variance
- But in practice??
- Before we saw how increasing the feature space can increase the complexity of the learned estimator:

$$\mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{F}_3 \subset \ldots$$

degree 1 poly　degree 2　3　4.　- - -

$$\widehat{f}_{\mathcal{D}}^{(k)} = \arg\min_{f \in \mathcal{F}_k} \frac{1}{|\mathcal{D}|} \sum_{(x_i, y_i) \in \mathcal{D}} (y_i - f(x_i))^2$$

Complexity grows as k grows

Training set error as a function of model complexity

$$\mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{F}_3 \subset \ldots$$

$$\widehat{f}_{\mathcal{D}}^{(k)} = \arg\min_{f \in \mathcal{F}_k} \frac{1}{|\mathcal{D}|} \sum_{(x_i, y_i) \in \mathcal{D}} (y_i - f(x_i))^2$$

TRAIN error:

$$\mathcal{D} \overset{i.i.d.}{\sim} P_{XY}$$

$$\underbrace{\frac{1}{|\mathcal{D}|} \sum_{(x_i, y_i) \in \mathcal{D}} (y_i - \widehat{f}_{\mathcal{D}}^{(k)}(x_i))^2}$$

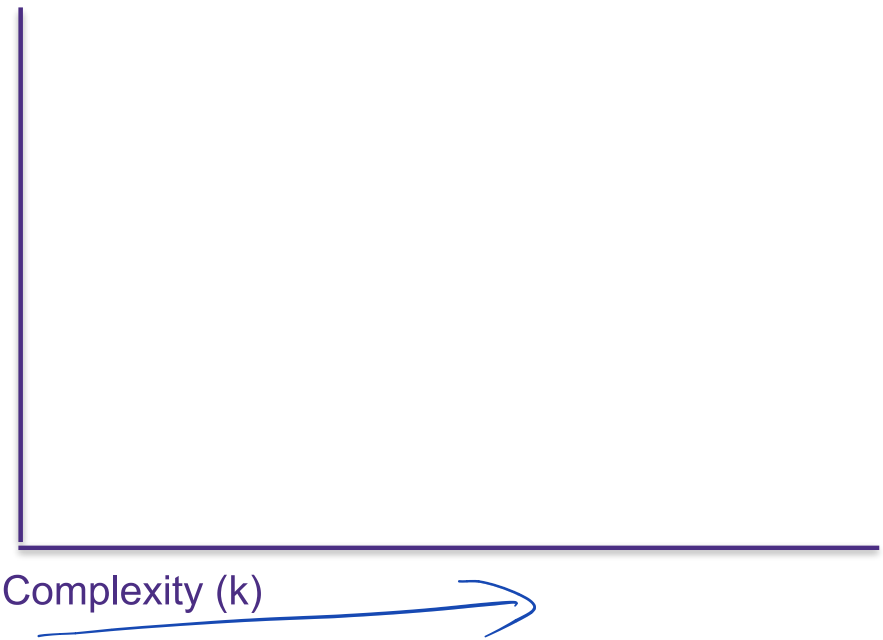
TRUE error:

$$\underline{\mathbb{E}_{XY}[(Y - \widehat{f}_{\mathcal{D}}^{(k)}}(X))^2]$$

Training set error as a function of model complexity

$$\mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{F}_3 \subset \dots$$

$$\widehat{f}_{\mathcal{D}}^{(k)} = \arg\min_{f \in \mathcal{F}_k} \frac{1}{|\mathcal{D}|} \sum_{(x_i, y_i) \in \mathcal{D}} (y_i - f(x_i))^2$$



TRAIN error:

$$\mathcal{D} \overset{i.i.d.}{\sim} P_{XY}$$

$$\frac{1}{|\mathcal{D}|} \sum_{(x_i, y_i) \in \mathcal{D}} (y_i - \widehat{f}_{\mathcal{D}}^{(k)}(x_i))^2$$

TRUE error:

$$\mathbb{E}_{XY}[(Y - \widehat{f}_{\mathcal{D}}^{(k)}(X))^2]$$

TEST error:

$$\mathcal{T} \overset{i.i.d.}{\sim} P_{XY}$$

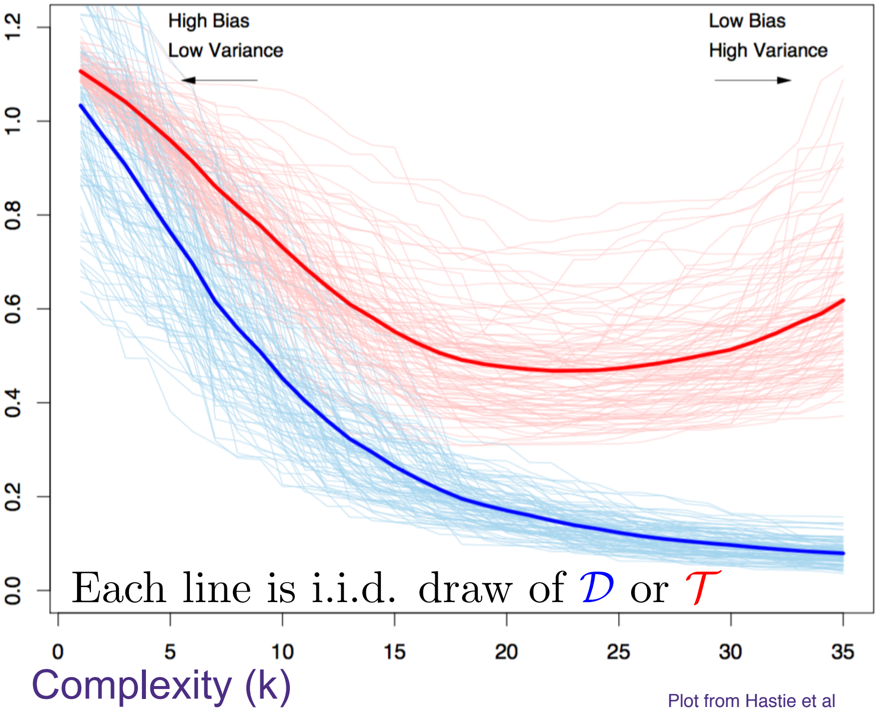
$$\frac{1}{|\mathcal{T}|} \sum_{(x_i, y_i) \in \mathcal{T}} (y_i - \widehat{f}_{\mathcal{D}}^{(k)}(x_i))^2$$

$$\boxed{\text{Important: } \mathcal{D} \cap \mathcal{T} = \emptyset}$$

Training set error as a function of model complexity

$$\mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{F}_3 \subset \dots$$

$$\widehat{f}_{\mathcal{D}}^{(k)} = \arg\min_{f \in \mathcal{F}_k} \frac{1}{|\mathcal{D}|} \sum_{(x_i,y_i)\in\mathcal{D}} (y_i - f(x_i))^2$$



Plot from Hastie et al

TRAIN error:

$$\mathcal{D} \overset{i.i.d.}{\sim} P_{XY}$$

$$\frac{1}{|\mathcal{D}|} \sum_{(x_i,y_i)\in\mathcal{D}} (y_i - \widehat{f}_{\mathcal{D}}^{(k)}(x_i))^2$$

TRUE error:

$$\mathbb{E}_{XY}[(Y - \widehat{f}_{\mathcal{D}}^{(k)}(X))^2]$$

= irr. err. + bias^2 + variance

TEST error:

$$\mathcal{T} \overset{i.i.d.}{\sim} P_{XY}$$

$$\mathbb{E}\left[\frac{1}{|\mathcal{T}|} \sum_{(x_i,y_i)\in\mathcal{T}} (y_i - \widehat{f}_{\mathcal{D}}^{(k)}(x_i))^2\right]$$

$$\boxed{\text{Important: } \mathcal{D} \cap \mathcal{T} = \emptyset}$$

Training set error as a function of model complexity

$$\mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{F}_3 \subset \ldots$$

$$\widehat{f}_{\mathcal{D}}^{(k)} = \arg\min_{f \in \mathcal{F}_k} \frac{1}{|\mathcal{D}|} \sum_{(x_i,y_i) \in \mathcal{D}} (y_i - f(x_i))^2$$

TRAIN error is optimistically biased because it is evaluated on the data it trained on. TEST error is unbiased only if T is never used to train the model or even pick the complexity k.

TRAIN error:

$$\mathcal{D} \overset{i.i.d.}{\sim} P_{XY}$$

$$\frac{1}{|\mathcal{D}|} \sum_{(x_i,y_i) \in \mathcal{D}} (y_i - \widehat{f}_{\mathcal{D}}^{(k)}(x_i))^2$$

TRUE error:

$$\mathbb{E}_{XY}[(Y - \widehat{f}_{\mathcal{D}}^{(k)}(X))^2]$$

TEST error:

$$\mathcal{T} \overset{i.i.d.}{\sim} P_{XY}$$

$$\frac{1}{|\mathcal{T}|} \sum_{(x_i,y_i) \in \mathcal{T}} (y_i - \widehat{f}_{\mathcal{D}}^{(k)}(x_i))^2$$

$$\boxed{\text{Important: } \mathcal{D} \cap \mathcal{T} = \emptyset}$$

Test set error

> Given a dataset, randomly split it into two parts:
  - Training data: $\mathcal{D}$
  - Test data: $\mathcal{T}$

$$\boxed{\text{Important: } \mathcal{D} \cap \mathcal{T} = \emptyset}$$

> Use training data to learn predictor
  - e.g., $\frac{1}{|\mathcal{D}|} \sum_{(x_i, y_i) \in \mathcal{D}} (y_i - \widehat{f}_{\mathcal{D}}^{(k)}(x_i))^2$
  - use training data to pick complexity k

> Use test data ~~to report predicted performance~~

$$\frac{1}{|\mathcal{T}|} \sum_{(x_i, y_i) \in \mathcal{T}} (y_i - \widehat{f}_{\mathcal{D}}^{(k)}(x_i))^2$$

How many points do I use for training/testing?

- **Very hard question to answer!**
  - **Too few training points, learned model is bad**
  - **Too few test points, you never know if you reached a good solution**
- **Bounds, such as Hoeffding's inequality can help:**

$$P(|\ \widehat{\theta} - \theta^*\ | \geq \epsilon) \quad \leq \quad 2e^{-2N\epsilon^2}$$

- **More on this later the quarter, but still hard to answer**
- **Typically:**
  - **If you have a reasonable amount of data 90/10 splits are common**
  - **If you have little data, then you need to get fancy (e.g., bootstrapping)**